Entropy **2015**, *17*, 1-x; doi:10.3390/——

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

A New Robust Regression Method Based on Minimization of Geodesic Distances on a Probabilistic Manifold: Application to Power Laws †

Geert Verdoolaege [1,2,*]

[1] Department of Applied Physics, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium; E-Mail: geert.verdoolaege@ugent.be; Tel.: +32-9-264-95-91; Fax: +32-9-264-41-98
[2] Laboratory for Plasma Physics—Royal Military Academy (LPP-ERM/KMS), Avenue de la Renaissancelaan 30, B-1000 Brussels, Belgium

† This paper is an extended version of our paper published in the 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), 21–26 September 2014, Amboise, France.

Academic Editors: Frédéric Barbaresco and Ali Mohammad-Djafari

Received: 3 April 2015 / Accepted: 25 June 2015 / Published: xx

---

Abstract: In regression analysis for deriving scaling laws that occur in various scientific disciplines, usually standard regression methods have been applied, of which ordinary least squares (OLS) is the most popular. In many situations, the assumptions underlying OLS are not fulfilled, and several other approaches have been proposed. However, most techniques address only part of the shortcomings of OLS. We here discuss a new and more general regression method, which we call geodesic least squares regression (GLS). The method is based on minimization of the Rao geodesic distance on a probabilistic manifold. For the case of a power law, we demonstrate the robustness of the method on synthetic data in the presence of significant uncertainty on both the data and the regression model. We then show good performance of the method in an application to a scaling law in magnetic confinement fusion.

Keywords: regression analysis; information geometry; geodesic distance; scaling laws; nuclear fusion

---

1. Introduction

Regression analysis is an essential instrument for data analysis in numerous branches of science. It is used for investigating deterministic relations between variables, for model building and for prediction by extrapolation to a previously unseen range of the involved variables. In this paper, we focus on regression analysis applied to the estimation of scaling laws. In various scientific disciplines, such as astronomy, biology, ecology and geology, scaling laws are used to characterize the underlying mechanisms at work in the respective complex systems under study. In general, a scaling law describes how a quantity of interest y scales when changing other quantities $x_1, x_2, \ldots, x_P$, on which it depends. Scaling laws are often expressed in terms of a power law:

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \ldots x_P^{\beta_P}. \tag{1}$$

A crucial property of such a power law is scale-invariance, *i.e.*, when multiplying any of the variables x_i by a constant a, the power law in Equation (1) essentially remains the same, being multiplied only by a constant a^{β_i}.

In nuclear fusion experiments based on magnetic confinement of a hot hydrogen plasma, scaling laws are crucial for predicting the performance of future fusion reactors, which will have a larger size, magnetic field, plasma density, *etc*., compared to present-day experimental devices [1]. These scaling laws can be estimated on the basis of datasets from multiple fusion devices, spanning a significant part of the parameter space. Ordinary least squares regression (OLS) combined with frequentist theory is the statistical workhorse that is employed for this purpose in the vast majority of cases. However, often, there is considerable uncertainty in the experimental data, including the predictor variables, and in the model equations (regression model). However, OLS only deals with uncertainty on the response variables and does not cover additional complications, including atypical observations (outliers), heteroscedasticity, correlations, non-Gaussian distributions, *etc*. As such, OLS regression is often unsuitable for deriving scaling laws [2,3], and many scientific fields could benefit greatly from a unified regression methodology that is flexible and robust and yet relatively simple to implement.

In order to be able to handle the complications mentioned above, we have developed a new regression method, called geodesic least squares regression (GLS). It is based on minimization of the Rao geodesic distance between probability distributions on a manifold equipped with the Fisher metric. In this paper, we briefly introduce the method by means of a simple example involving a power law and Gaussian noise. We show the good performance of the method on synthetic data, introducing outliers in the first case and studying the effect of a logarithmic transformation of the data in the second case. Finally, we present an application to the important scaling concerning the power threshold for the transition into the high confinement regime (H-mode) in nuclear fusion experiments based on magnetic plasma confinement. The details of the quantities involved in this scaling, their experimental determination and the underlying physics are not important for the purpose of this paper. Rather, we here aim at showing the performance of GLS on a challenging and heterogeneous real-life dataset.

The paper is structured as follows. The method of geodesic least squares regression is described in Section 2, including a short discussion on calculating geodesic distances on a Gaussian probabilistic manifold, within the framework of information geometry. The next section, Section 3, briefly introduces the database that is used in the subsequent regression experiments, in relation to the scaling law for the

H-mode power threshold in fusion plasmas. The experiments involving synthetic data are described in Section 4, while the real power threshold scaling is derived in Section 5. Section 6 concludes the paper and contains an outlook towards future work related to the methodology.

2. Geodesic Least Squares Regression

We start by describing the GLS methodology, which was already introduced in [4,5], but here, we go into some more detail. We describe a specific form of GLS regression, and it should be stressed that various aspects can be generalized, as will be noted accordingly. Furthermore, several elements on which GLS is based are also found in other regression techniques. The strength of GLS regression is that it integrates many of these aspects in an elegant way, resulting in a method that is very general, flexible and robust. From one point of view, GLS is similar to a class of parameter estimation methods that are collectively referred to in the statistics community as minimum distance estimation, in that GLS minimizes a distance between a parametric model distribution of the data and an empirical distribution [6]. We use the Rao geodesic distance (GD) as a similarity measure, which has the advantage that it offers an intuitive geometric interpretation. In addition, there are similarities between GLS and the generalized linear model [7].

We will consider the case of regression with multiple predictor variables (regressors) and a single response variable. For this case, we will show that GLS regression can be regarded as a generalization of OLS. However, GLS takes place on a probabilistic manifold, whereas classic OLS operates in a flat Euclidean space. Indeed, OLS is based on minimizing the difference, *i.e.*, the Euclidean distance, between the predicted and measured values of the response variable. Likewise, GLS is based on minimizing the GD between distributions on the probabilistic manifold. Therefore, we start by briefly introducing some concepts from information geometry related to distance calculation.

2.1. Distance in Information Geometry

In information geometry, a parametric family of probability densities is interpreted as a Riemannian differentiable manifold [8]. Each point on the manifold corresponds to a specific probability density function (pdf) within the family, and the family parameters represent a coordinate system on the manifold. The Fisher information (covariance of the score) provides a unique metric tensor. For a probability model $p(\{x_m\}|\{\theta^k\})$ [9] describing a set $\{x_m\}$ of M variables ($m = 1,\ldots,M$), parameterized by a set $\{\theta^k\}$ of P parameters ($k = 1,\ldots,P$), the entries g_{ij} of the Fisher information matrix are given by (no summation):

$$g_{ij}\big(\{\theta^k\}\big) = \mathbb{E}\left[\frac{\partial}{\partial\theta^i}\ln p\big(\{x_m\}|\{\theta^k\}\big)\frac{\partial}{\partial\theta^j}\ln p\big(\{x_m\}|\{\theta^k\}\big)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^i\partial\theta^j}\ln p\big(\{x_m\}|\{\theta^k\}\big)\right], \qquad i,j,k = 1,\ldots,P.$$

The metric provides the basis for distance measurement between pdfs. Specifically, a geodesic curve locally minimizes the distance between two points on the manifold equipped with that metric. Through calculus of variations, it can be shown that a geodesic is the solution of the following system of

nonlinear second-order ordinary differential equations, known in the language of variational analysis as Euler–Lagrange equations [10] and in the present context as geodesic equations:

$$\ddot{\theta}^r(t) + \sum_{i,j=1}^{P} \Gamma^r{}_{ij}\, \dot{\theta}^i(t)\dot{\theta}^j(t) = 0, \qquad r = 1, \ldots, P. \tag{2}$$

Here, the θ^i have been parameterized along the geodesic by t and $\Gamma^r{}_{ij}$ are the Christoffel symbols of the second kind, defined through the metric as:

$$\Gamma^k{}_{ij} = \frac{1}{2} \sum_r g^{kr} \left(\frac{\partial g_{jr}}{\partial \theta^i} + \frac{\partial g_{ir}}{\partial \theta^j} - \frac{\partial g_{ij}}{\partial \theta^r} \right),$$

where g^{ij} denotes the components of the inverse metric. The boundary value problem Equation (2) needs to be solved assuming the known values of the coordinates at the boundary points of the geodesic.

From the metric and the solution of the geodesic equations, the length L_g of the geodesic curve between two distributions with parameter sets $\{\theta_1^i\}$ and $\{\theta_2^i\}$, *i.e.*, the geodesic distance between these distributions, may be locally calculated as follows (assuming t runs from zero to one):

$$L_g = \int_{\{\theta_1^i\}}^{\{\theta_2^i\}} \mathrm{d}s = \int_0^1 \left(\sum_{i,j} g_{ij} \dot{\theta}^i \dot{\theta}^j \right)^{1/2} \mathrm{d}t, \tag{3}$$

where s represents the arc length. In the framework of information geometry, the geodesic distance based on the Fisher metric is often referred to as the Rao geodesic distance (GD).

Coming back to Equations (2) and (3), it should be noted that closed-form expressions for the GD are rarely available. On the other hand, provided the Fisher metric can be calculated relatively easily, the framework of information geometry is very useful, since straightforward approximations of the geodesic curves can be found in a geometrically intuitive way [11]. This intuitive approach by means of geometry is an important and attractive aspect of the theory, as it provides enhanced insight into various concepts and algorithms in probability theory and statistics [12]. This is also the case for GLS, as will be demonstrated below. Furthermore, as far as the GD is concerned, visualization of geodesics may guide controlled approximations to the geodesic paths and geodesic distances [11].

Besides the attractive feature of providing intuitive geometrical insight into problems involving similarity measurement between probability distributions, the GD has several more advantages over other similarity measures for distributions. First, it is a distance measure (a metric) in the strict sense of the word. As a result, it is symmetric in its arguments, a desirable property for measuring the similarity between two given states of information in terms of probability distributions. In addition, it obeys the triangle inequality, yielding various practical advantages, for instance in the field of data retrieval from large databases [13]. Furthermore, closed-form expressions may be available for the GD, or its approximation, for various families of distributions where no such analytic form has been found in the case of, for instance, the Kullback–Leibler divergence (KLD) [11]. Finally, there is considerable experimental evidence suggesting that the GD in general is a more effective similarity measure between distributions than the KLD (see [11] and the references therein). We note that for distributions that lie infinitesimally close on the probabilistic manifold, it can be proven that the Kullback–Leibler divergence equals half of the squared geodesic distance between the distributions (see, e.g., [14]). Hence, in such a

case, the KLD and GD yield similar results, but in general, they are quite different measures of similarity between distributions.

2.2. Geodesics for the Univariate Normal Distribution

In this paper, we discuss applications that are based on a univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$, parameterized by its mean μ and standard deviation σ. In this case, an analytic expression for the Fisher–Rao metric is available. It turns out to be the familiar Poincaré metric, which, when applied to a half-plane, is a well-known model for hyperbolic geometry that has constant negative scalar curvature. The line-element is given by [15,16]:

$$\mathrm{d}s^2 = \frac{\mathrm{d}\mu^2}{\sigma^2} + 2\frac{\mathrm{d}\sigma^2}{\sigma^2}. \tag{4}$$

As an illustration, the Poincaré half-plane is pictured in Figure 1a, together with two geodesics between, on the one hand, the points $p_1 = \mathcal{N}(4, 1.2^2)$ and $p_2 = \mathcal{N}(16, 1.5^2)$ and, on the other hand, $p_3 = \mathcal{N}(4, 4.0^2)$ and $p_4 = \mathcal{N}(16, 5.0^2)$. The corresponding normal density functions are drawn in Figure 1b, as well as a number of densities associated with some intermediate points on each geodesic. As a further illustration, Figure 1c shows one blade of a particular pseudosphere, namely the tractroid, which is locally isometric to the Poincaré half-plane and the univariate normal manifold for $\sigma > 1$, with periodicity in μ. In order to better visualize the geodesics, a rescaled version of the tractroid is shown in Figure 1d. This surface has a longer period in the μ-direction. However, it should be kept in mind that only the visualization in Figure 1c can be used to measure absolute distances on the surface, for in Figure 1d, the pictured geodesics are no longer the shortest curves between the points in question. It is clear that the geodesics on the Gaussian manifold are different from straight lines in the Euclidean space, wherein the manifold has been immersed. The shape of the geodesics can be made intuitively clear by noting that they always pass through a region of increased standard deviation relative to that of the boundary points. This provides the shortest route, as can be seen from the line element Equation (4). Interestingly, similar arguments will be shown to enable a deeper insight into the operation of GLS regression. We further note that various alternative models exist to visualize hyperbolic geometry; see, e.g., [17].

A closed-form expression is available for the GD on the normal manifold, permitting fast evaluation. Indeed, for two univariate normal distributions $p_1(x|\mu_1, \sigma_1^2)$ and $p_2(x|\mu_2, \sigma_2^2)$, the GD is given by [16]:

$$\mathrm{GD}(p_1, p_2) = \sqrt{2}\ln\frac{1+\delta}{1-\delta} = 2\sqrt{2}\tanh^{-1}\delta, \qquad \delta \equiv \left[\frac{(\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2 + 2(\sigma_1+\sigma_2)^2}\right]^{1/2}. \tag{5}$$

Furthermore, since the injectivity radius of the hyperbolic plane is infinite, the geodesics are globally length-minimizing [10].

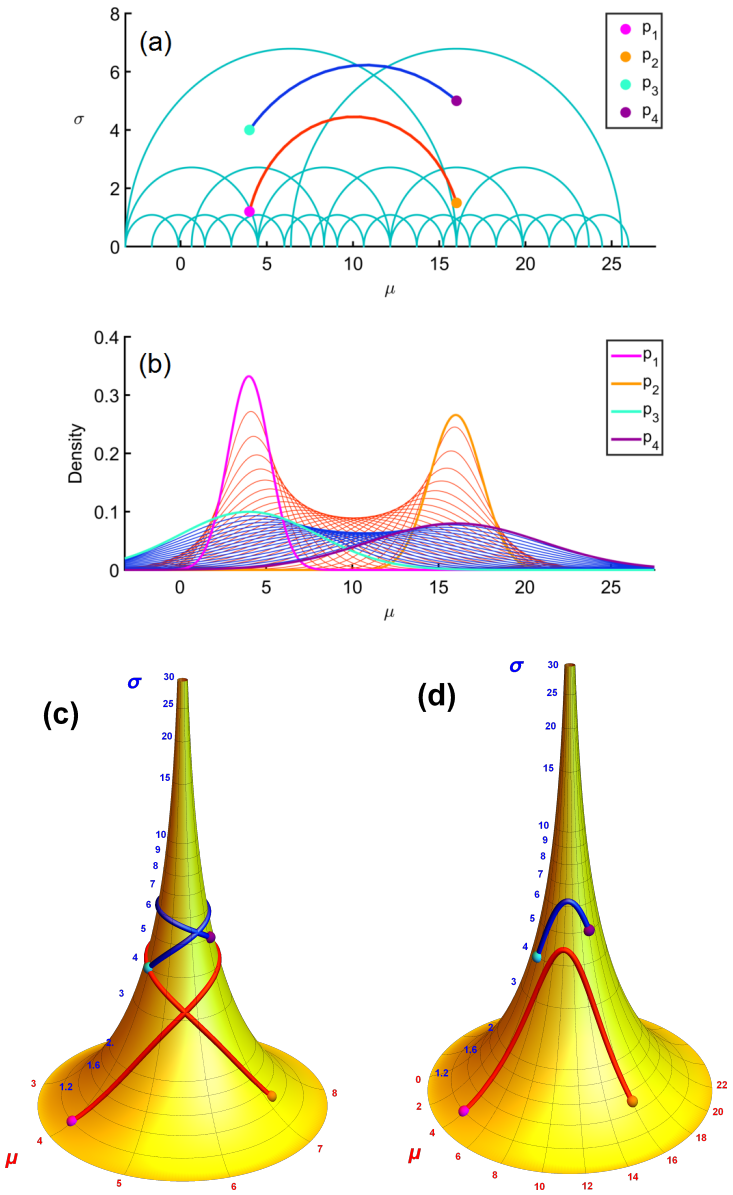

Figure 1. (**a**) Illustration of the Poincaré half-plane with several half-circle geodesics in the background, together with the geodesic between the points p_1 and p_2 and between p_3 and p_4, defined in the main text. (**b**) Probability densities corresponding to the points p_1, p_2, p_3 and p_4 indicated in (a). The densities associated with some intermediate points on the geodesics between p_1 and p_2 and between p_3 and p_4 are also drawn. (**c**) Rendering of one blade of the tractroid, again with the two geodesics superimposed. The parallels of the tractroid are lines of constant standard deviation σ, while the meridians (the tractrices) are lines of constant mean μ. This representation of the normal manifold is periodic in the μ-direction, and a rescaled version (longer period along μ) is shown in (**d**).

2.3. Geodesic Least Squares Methodology

GLS starts from the premise that the probability distribution underlying experimental measurements is the fundamental object resulting from the measurement. As such, GLS does not perform regression based on data points in a Euclidean space, but rather operates on probability distributions lying on

a probabilistic manifold. This introduces additional flexibility that renders the method robust in the presence of large uncertainties, as will be demonstrated in the experiments.

Briefly, the idea is to consider two different proposals for the distribution of the response (dependent) variable y, conditional on the predictor variables. On the one hand, there is the distribution that one would expect if all assumptions were correct regarding the deterministic component of the regression model (regression function) and the stochastic component. We call this the modeled distribution. On the other hand, we try to capture the conditional distribution of y by relying less on the model assumptions, but directly on the measurements of y. For this, we will use the term observed distribution. It is in this sense that GLS is similar to minimum distance estimation (MDE), where the Hellinger distance is a popular similarity measure [18]. This was first applied to regression in [19], but there are several differences with GLS. First and foremost, GLS calculates the geodesic distance between each individual pair of modeled and observed distributions of the response variable, corresponding to an individual measurement point. As such, each individual data point acquires the status of a probability distribution in its own right. Consequently, GLS performs regression between probability distributions on a probabilistic manifold. In contrast, MDE usually considers a distance between a kernel density estimate of the distribution of residuals, on the one hand, and the parametric model, on the other hand, based on the entire data sample. Secondly, we explicitly model all parameters of the modeled distribution, which is similar to the ideas behind the link function in the generalized linear model (GLM) [7]. In the present work, this will be accomplished by explicitly modeling both the mean and standard deviation of the Gaussian modeled distribution. Additionally, a final difference is that we use the Rao geodesic distance as a similarity measure.

As a simple example that we will use also in the experiments, consider a linear relation $\eta = \beta\xi$ between a single predictor variable ξ and a response variable η, with β a constant. In accordance with the discussion above, we explicitly wish to allow for the challenging case of uncertainty on the predictor variable ξ. Therefore, we assume that, in reality, N samples of a stochastic (noisy) variable x are observed, together with N samples of a stochastic response variable y. We take the simple case of normally distributed (Gaussian) noise:

$$\begin{aligned} y &= \eta + \epsilon_y = \beta\xi + \epsilon_y, & \epsilon_y &\sim \mathcal{N}\left(0, \sigma_y^2\right), \\ x &= \xi + \epsilon_x, & \epsilon_x &\sim \mathcal{N}\left(0, \sigma_x^2\right). \end{aligned} \tag{6}$$

The observations x_n ($n = 1, \ldots, N$) are taken as mutually independent and so are the y_n. σ_x and σ_y are assumed to be known, and in this example, they are taken constant for all measurements, *i.e.*, we have homoscedasticity. However, we will also consider heteroscedasticity later on. According to the regression model, conditionally on x_n, each measurement y_n is drawn from a normal distribution:

$$p_{\mathrm{mod}}(y|x_n) = \mathcal{N}\left(\beta x_n, \sigma_{\mathrm{mod}}^2\right), \quad \text{where} \quad \sigma_{\mathrm{mod}}^2 \equiv \sigma_y^2 + \beta^2\sigma_x^2, \tag{7}$$

with the subscript "mod" referring to the modeled distribution. In our simple example, Equation (7) follows from standard Gaussian error propagation rules. However, for nonlinear regression laws, the conditional distribution for y has to be obtained by marginalizing the unknown true values ξ_n. Nevertheless, the Gaussian error propagation laws may be used in the nonlinear case as well, to approximate the conditional distribution $p(y|x_n)$ by a normal distribution, as will be shown in the experiments.

We next choose a specific form of the observed distribution corresponding to each realization of the variable y, conditional on the observations, *i.e.*, $p_{\text{obs}}(y|y_n)$. In this example, we take again the normal distribution, but centered on each data point: $\mathcal{N}(y_n, \sigma^2_{\text{obs}})$, where σ_{obs} is to be estimated from the data. In the context of the GLM, this is known as the saturated model. The extra parameter σ_{obs} gives the method added flexibility, since it is not *a priori* required to equal σ_{mod}. As a result, GLS is less sensitive to incorrect model assumptions. Note that in this example, we have chosen the observed distribution from the same model (Gaussian) as the modeled distribution. Furthermore, σ_{mod} is taken as a fixed value for all measurements and so is σ_{obs}. These assumptions can of course be relaxed, leading to a more general method. However, the transition from OLS to GLS is best explained by means of a Gaussian observed distribution, which, in addition, offers computational advantages, since the expression for the GD has a closed form; see Equation (5).

GLS now proceeds by minimizing the total GD between, on the one hand, the joint observed distribution of the N realizations of the variable y and, on the other hand, the joint modeled distribution. Thanks to the independence assumption in this example, we can write this in terms of products of the corresponding marginal distributions:

$$\begin{aligned}\hat{\beta} &= \underset{\beta\in\mathbb{R}}{\operatorname{argmin}}\, \mathrm{GD}\left[\prod_{n=1}^{N} p_{\text{obs}}(y|y_n), \prod_{n=1}^{N} p_{\text{mod}}(y|x_n)\right] \\ &= \underset{\beta\in\mathbb{R}}{\operatorname{argmin}} \sum_{n=1}^{N} \mathrm{GD}^2\left[p_{\text{obs}}(y|y_n), p_{\text{mod}}(y|x_n)\right]. \end{aligned} \tag{8}$$

The last equality entails a considerable simplification, owing to the property that the squared GD between products of distributions can be written as the sum of squared GDs between the corresponding factors [16]. Hence, the optimization procedure involves matching not only y_n with bx_n, but also σ^2_{obs} with $\sigma_y^2 + \beta^2\sigma_x^2$. Note that the parameter β occurs both in the mean and the variance of the modeled distribution. Incidentally, forcing $\sigma^2_{\text{obs}} \equiv \sigma_y^2 + \beta^2\sigma_x^2$ would take us back to standard maximum likelihood estimation, for the Rao GD between the two Gaussians p_{obs} and p_{mod} with means y_n and bx_n, respectively, but with identical standard deviations (fixed along the geodesic path), is precisely the Mahalanobis distance [20]:

$$\mathrm{GD}(p_{\text{obs}}, p_{\text{mod}}) = \frac{|y_n - bx_n|}{\sigma_y^2 + \beta^2\sigma_x^2}, \qquad \text{if } \sigma^2_{\text{obs}} \equiv \sigma_y^2 + \beta^2\sigma_x^2.$$

We note that the GLS scheme addresses many of the difficulties with classic OLS regression. First, GLS explicitly allows uncertainty on the predictor variables, and it is not restricted to normal or symmetric noise distributions, nor does it necessarily assume homoscedasticity. In addition, correlations among variables and among observations can be built into the stochastic component of the regression model. Furthermore, GLS can operate with any (nonlinear) regression function. Moreover, it will be shown in the experiments that GLS is relatively insensitive to uncertainties in both the stochastic and deterministic components of the regression model. The same quality renders the method also robust against outliers.

In the experiments below, we employed a classic active-set algorithm to carry out the optimization [21]. Furthermore, presently, the GLS method does not directly offer confidence (or

credible) intervals on the estimated quantities. Future work will address this issue in more detail, but for now, error estimates were derived by Monte Carlo sampling in the case of the numerical simulations (Section 4) and by bootstrapping in the case of the real data (Section 5) [22]. The bootstrapping involved creating, from the measured dataset, a large number of artificial datasets of the same size, by resampling with replacement. The regression analysis was then carried out on each of the datasets, and the mean and standard deviation, over all datasets, of each estimated regression parameter and of the predicted quantities were used as estimates of the parameter or prediction value and its error bar, respectively. This scheme typically results in rather conservative error bars, which could possibly be narrowed down using more sophisticated methods.

3. The L-H Power Threshold and Database

We now provide some background information regarding the main regression application that will be treated in the experiments with synthetic and real data. It concerns one of the most important scaling relations in fusion science based on magnetic plasma confinement, related to the threshold P_{thr} for the heating power that is required for the plasma to make the transition into a desired regime of high energy confinement (H-mode) in the next-step fusion device ITER (International Thermonuclear Experimental Reactor) [1,23,24]. To a good approximation, this so-called L-H (or H-mode) power threshold depends on the electron density in the plasma $\bar{n}_{\mathrm{e}}$ (in $10^{20}\ \mathrm{m}^{-3}$), the main magnetic field B_{t} (in tesla (T)) and the total surface area S of the confined plasma (in m^2). This is usually expressed by means of the following scaling relation:

$$P_{\mathrm{thr}} = \beta_0\, \bar{n}_{\mathrm{e}}^{\beta_1} B_{\mathrm{t}}^{\beta_2} S^{\beta_3}. \tag{9}$$

To estimate the coefficients in this relation, we employed data from eight fusion devices of the tokamak type (ASDEX, AUG, CMOD, DIII-D, JET, JFT-2M, JT-60U, PBXM) in the International Tokamak Physics Activity (ITPA) multi-machine database for the L-H power threshold (subset IAEA02 [23,25–27]). This yields 616 measurement sets of power, density, magnetic field and surface area, each set obtained during a brief time of plasma operation under stationary conditions in one of the eight devices involved in the study [28].

The ITPA database contains some information regarding the error bars on the measurements, specifically relative errors expressed as percentages. This is important for our purposes, because we need the error bars to calculate σ_{mod}. Unfortunately, the error estimates are not available in some cases, and if they are, the precise definition of the error bars is not always clear. Usually, an error bar in the database represents an estimate by the experimentalist of the typical range in which the "true" quantity can be expected to lie, where the uncertainty is assumed to be caused by both stochastic and systematic effects. Moreover, it is difficult to assess the probability that is covered by the stochastic component of the errors mentioned in the database. Since a detailed investigation of the uncertainty of the threshold data is beyond the scope of the present paper, we will assume that the error bars pertain to a stochastic uncertainty corresponding to a single standard deviation of a Gaussian distribution. For some derived quantities, the error bars had to be calculated from the uncertainty on more fundamental measurements. In those cases, we employed Gaussian error propagation rules to estimate the standard deviation on the

derived quantities. For the case of the global H-mode confinement database, this strategy has been shown to provide reasonable information on the actual error bars [29].

It is important to mention that the main source of uncertainty in the data used for power threshold scaling, when compared to the predictions of a simple power law regression model, is not expected to be due to the measurement uncertainty on the individual variables. There are far larger sources owing to the variability of the experiments that produced the data. To estimate the variability of each of the physical quantities with respect to the scaling law, we performed the following calculation. First, the nonlinear scaling law was estimated using OLS, as explained in Section 5.2. Then, for a specific variable z (one of the predictor variables or the dependent variable) and for each data point, the relative difference was computed between the z-value of the data point itself and the z-value of the projection of the data point on the hypersurface given by the scaling law, keeping the values of the other variables fixed. This difference can be interpreted as the deviation of the point from the theoretical scaling law, assuming the deviation is solely due to the variability of the variable z. Finally, the standard deviation of these relative differences was taken, and the procedure was repeated for every predictor variable and the dependent variable. The resulting standard deviations can be interpreted as upper bounds of the relative variability of each of the variables around their 'theoretical' values given by the scaling law. This way, for $\bar{n}_{\mathrm{e}}$, B_{t}, S and P_{thr}, we obtained 39%, 31%, 28% and 38%, respectively. These levels are clearly much larger than the relative uncertainties due to measurement error alone. Indeed, the typical measurement error bars quoted in the ITPA database, on average, over all devices, are estimated at 4% for $\bar{n}_{\mathrm{e}}$, 1% for B_{t}, 3% for S and 15% for P_{thr} [25,26].

4. Numerical Simulations

We next demonstrate some of the potential of the GLS regression scheme by means of a number of experiments with synthetically-generated data. We treat two particular cases of deviation from the model according to which the data were created and show that, in comparison with a number of standard regression techniques, GLS yields the most accurate results across all experiments. The first case concerns the effect of outliers, while in the second case, the influence of a logarithmic transformation is studied. In each case, we start with a very simple experiment that is easily reproduced, using a single predictor variable, providing some intuitive feeling regarding the performance of the method. We then proceed to a more elaborate test, still based on partly synthetic data, but using a regression challenge similar to that used in the real-world experiment for scaling of the L-H power threshold in fusion plasmas, presented in Section 5.

4.1. Effect of Outliers

The robustness of minimum distance estimators to outliers in the data was noted in the classic literature of minimum distance estimation [18]. We now show that this is a quality also enjoyed by GLS regression.

4.1.1. Single Predictor Variable

We first concentrate on estimating the slope of a regression line with a single predictor variable. To this end, a dataset was generated consisting of ten points labeled by coordinates ξ_n and η_n ($n = 1, \ldots, 10$), with the ξ_n chosen unevenly between zero and 50 and $\eta_n = \beta\xi_n$, taking $\beta = 3$. Then, Gaussian noise was added to all coordinates according to Equation (6), with $\sigma_y = 2.0$ and $\sigma_x = 0.5$, resulting in values x_n for the predictor variable and y_n for the response variable. Finally, one outlier was created by multiplying the value of y_k by a factor distributed uniformly between 1.5 and 2.5, with k chosen uniformly among the indices 8, 9 and 10.

We next estimated β by means of GLS and compared the estimates with those obtained by OLS, maximum *a posteriori* (MAP) using the model in Equation (7) for the likelihood and an uninformative prior [30], total least squares (TLS), which is a typical errors-in-variables technique [31], and a robust method (ROB) based on iteratively reweighted least squares (bisquare weighting) [32], included in the MATLAB Statistics toolbox [33]. It should be noted that MAP takes into account the error bars on the predictor variables. In all cases, we assumed knowledge of the values of σ_x and σ_y. In order to get an idea of the variability of the estimates, Monte Carlo sampling of the data-generating distributions was performed, and the estimation was carried out 100 times.

The results are given in Table 1, mentioning the sample average and standard deviation of the estimates $\hat{\beta}$ over the 100 runs for each of the methods. GLS is seen to perform very well and similar to the robust method ROB, but the other techniques yield considerably worse results. The average estimate of σ_{obs} was 5.43 with a standard deviation of 0.24. On the other hand, the modeled value of the standard deviation in the conditional distribution for y was $\sigma_{\text{mod}} = \sqrt{\sigma_y^2 + 9\sigma_x^2} = 2.5$. Hence, GLS succeeds in ignoring the outlier by increasing the estimated variability of the data. Put differently, the effect of the outlier is, in a sense, to increase the overall variability of the data, which GLS takes into account by increasing the observed standard deviation of the data (σ_{obs}) with respect to the standard deviation predicted by the model (σ_{mod}).

Table 1. Monte Carlo estimates of the mean and standard deviation for the slope parameter in linear regression with errors on both variables and one outlier. GLS, geodesic least squares regression; TLS, total least squares; ROB, robust method.

Original	**GLS**	**OLS**	**MAP**	**TLS**	**ROB**
$\beta = 3.00$	$\mathbf{3.031 \pm 0.035}$	3.68 ± 0.29	3.83 ± 0.36	4.6 ± 1.0	2.992 ± 0.041

As mentioned before, this result can also be understood in terms of the pseudosphere as a geometrical model for the normal distribution. To see this, we refer to Figure 2, where several sets of points (distributions) are drawn on a portion of the surface of the pseudosphere for one particular dataset generated as described above. First, the modeled distributions are plotted with their means $\hat{\beta}x_n$ (see Equation (7)) and standard deviations $\sigma_{\text{mod}} = 2.5$, using the average estimate $\hat{\beta} = 3.031$ obtained by GLS. These are the green points on the surface, and they lie on a parallel, since they all correspond to Gaussians with the same standard deviation σ_{mod}. In this particular dataset, the index of the outlier was $k = 10$, so the point $\hat{\beta}x_{10}$ is indicated individually. Obviously, according to the model, no outlier is

expected, so the modeled distribution corresponding to $k = 10$, which is the green point just mentioned, lies close to the other predicted points (distributions). Next, we plot the observed distributions with their means y_n and standard deviations σ_{obs} (for this dataset estimated at $\hat{\sigma}_{\mathrm{obs}} = 5.43$). These are the blue points, lying at a constant standard deviation σ_{obs}, which is higher than σ_{mod} ($5.43 > 2.5$). The outlier y_{10} can clearly be observed, and being an outlier, it lies relatively far away from the rest of the blue points (observed distributions). Now suppose that, like MAP, GLS would not be able to increase σ_{obs} relative to σ_{mod} in order to accommodate the outlier. Then, the observed distributions would have the same observed means (the measured values y_n), but they would have the standard deviation predicted by the model. Hence, they would lie on the parallel corresponding to σ_{mod}, just like the green points. We have plotted these fictitious distributions as the red points at the level of σ_{mod}, and they are labeled $\tilde{y}$. Again, the outlier (labeled $\tilde{y}_{10}$) can be seen, but it seems to lie further away from the other red points (the points $\tilde{y}_n$) compared to the actually observed situation, *i.e.*, the distance from y_{10} to the other y_n (blue points). At least this is the case when using (visually) the Euclidean distance in the embedding Euclidean space. We can verify that this is indeed so by using the proper geodesic distance on the surface: overall, the blue points lie closer together (including the outlier) than the red points. Now, in fact, GLS aims at minimizing the distance between each green point (modeled distribution) and its corresponding blue point (observed distribution), so as far as the outlier is concerned, we should really be looking at the geodesic between the point $(\hat{\beta}x_{10}, \sigma_{\mathrm{mod}})$ and the point $(y_{10}, \sigma_{\mathrm{obs}})$. The geodesic (labeled "Geo_1") between these points is also drawn on the surface, and again, we compare this to the fictitious situation, represented by the geodesic (labeled "Geo_2") between $(\hat{\beta}x_{10}, \sigma_{\mathrm{mod}})$ and $(\tilde{y}_{10}, \sigma_{\mathrm{mod}})$. Indeed, again, we see that the geodesic Geo_1 is shorter than Geo_2. Therefore, by increasing σ_{obs} relative to σ_{mod}, the outlier is not so much an outlier anymore, as measured on the pseudosphere!When calculating the GD, one finds 2.4 for Geo_1 and 2.8 for Geo_2. Therefore, GLS obtains a lower value of the objective function (sum of squared geodesic distances) if it increases σ_{obs} with respect to σ_{mod}. Of course, there is a limit to this: GLS cannot continue raising σ_{obs} indefinitely, trying to mitigate the distorting effect of the outlier, for then, the other points would get a too high observed standard deviation, which is not supported by the data. The image that we see in Figure 2 is the best compromise that GLS could find. In fact, we note that, in the case we suspect that y_{10} could be an outlier, it may very well be worthwhile to introduce two parameters to describe the observed standard deviation: one for the nine points that seem to follow the model and one to take care of the outlier. This would be a very straightforward extension of the method, and we explore this to some extent when using data from the ITPA database below. There, we assign a separate parameter to describe the observed standard deviation of all data coming from a specific tokamak, hence defining an individual parameter for each machine.

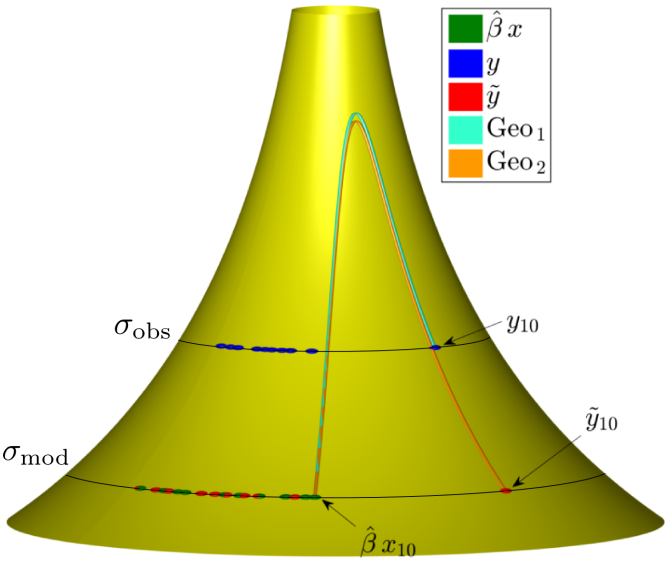

Figure 2. A portion of the pseudosphere together with the regression results on synthetic data with an outlier, as described in the main text.

4.1.2. Multiple Predictor Variables

In this experiment, a regression problem with multiple predictor variables and a power law is studied. The deterministic part of the regression model is based on the real-world problem for the L-H power threshold in fusion plasmas, which we are going to consider in Section 5. Furthermore, the values of the predictor variables are taken from the same international power threshold database, and values of the response variable are synthetically generated from this.

More specifically, the dataset for this experiment was created as follows. First, an artificial linear regression law was put forward for a variable η, depending on the predictor variables $\bar{n}_{\mathrm{e}}$, B_{t} and S, which were introduced in the context of the power threshold scaling law in Section 3 [34]. In particular, we generated a number of realizations of the variable η from the following prescription:

$$\eta = \beta_0 + \beta_1 \bar{n}_{\mathrm{e}} + \beta_2 B_{\mathrm{t}} + \beta_3 S. \tag{10}$$

This was considered as the "true" relation between the predictor and response variables, where, as mentioned above, the values of the predictor variables were chosen to be exactly those from the ITPA database, which are normally used in the real power threshold scaling law. A whole range of datasets was created using the following values of the coefficients β_0, β_1, β_2 and β_3:

$$\begin{aligned} \beta_0 &= 1, 1.1, \ldots, 20, \\ \beta_1, \beta_2, \beta_3 &= 0.1, 0.2, \ldots, 2. \end{aligned} \tag{11}$$

Thus, for each combination of values of β_0, β_1, β_2 and β_3, all 616 values of η were calculated according to Equation (10), based on the values of $\bar{n}_{\mathrm{e}}$, B_{t} and S from the ITPA database. The range of coefficient values in Equation (11) was chosen to be representative for the values that are typically obtained from a regression analysis on the true scaling law (see Section 5). The exception is β_0, for which the range was chosen of roughly the same order as $\eta - \beta_0$ (much smaller values of β_0 would not be estimable in comparison with $\eta - \beta_0$).

Next, Gaussian noise was added to both the predictor and response variables. The noise level was chosen according to the typical relative measurement errors in the ITPA database, *i.e.*, 4% for $\bar{n}_{\mathrm{e}}$, resulting in a variable x_1, 1% for B_{t} (variable x_2), 3% for S (variable x_3) and 15% for the dependent variable (variable y, which is P_{thr} in the real-world regression problem). It should be stressed that, in the light of our comments in Section 3 regarding the variability of the predictor variables, these are rather low noise levels. We further note that fixed relative noise levels lead to a different standard deviation for each measurement (heteroscedasticity).

Furthermore, in this experiment studying the effect of atypical observations, 10 outliers were created in each of the datasets. In particular, from the total of 616 points in each dataset, 10 points were randomly chosen, and the associated value of y was multiplied with a factor F, where F was distributed uniformly between 1.5 and 2.5. For each combination of coefficient values β_i ($i = 0, \ldots, 3$) taken from Equation (11), 10 datasets were realized, each time performing the sampling of noise and outliers.

Finally, the regression analysis was carried out for every dataset, and for each choice of the β_i, the obtained estimates $\hat{\beta}_i$ were defined as the average over the 10 data realizations. Next, histograms were created based on these averages for the estimated coefficients, specifically the normalized histograms of the relative difference $(\beta_i - \hat{\beta}_i)/\beta_i$ ($i = 0, \ldots, 3$), expressed as a percentage, between the true value β_i and the estimated value $\hat{\beta}_i$ of each regression parameter. The histograms of these percentage errors are shown in Figure 3. In order to avoid a cluttered figure, the results of OLS, MAP and GLS are plotted in one panel and those of TLS and ROB in another.

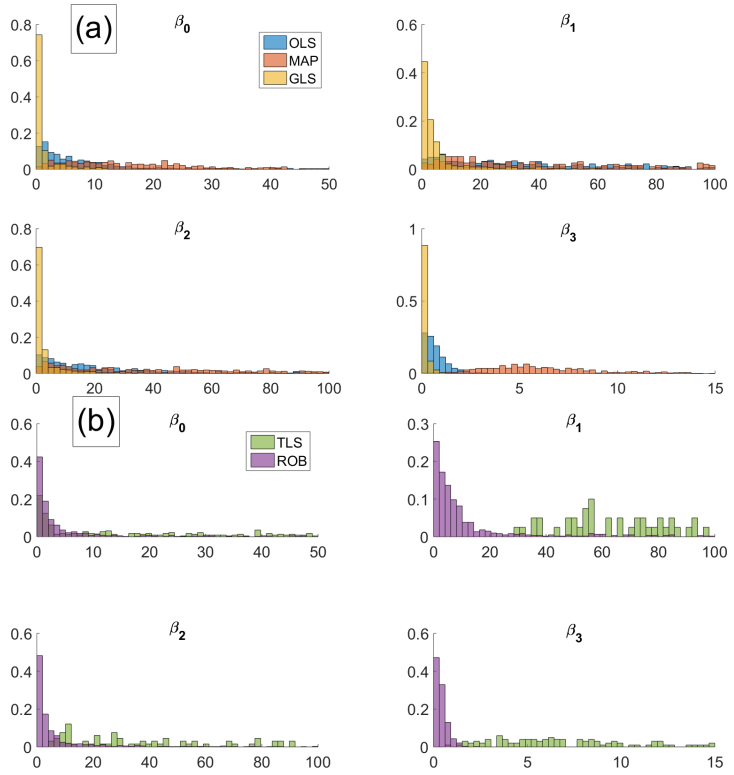

Figure 3. (**a**) Histograms of the relative error in estimating the regression coefficients β_i by means of OLS, MAP and GLS for a linear regression problem with outliers. Horizontal axes represent the error in percent and vertical axes probability, normalized to one. (**b**) Similar, for TLS and ROB.

From these histograms, it is clear that, for each parameter, GLS performs much better than OLS, MAP and TLS, with the latter failing completely. In case of GLS, the vast majority of relative errors is of the order of a few percent and certainly smaller than 20%. Overall, the most difficult to estimate parameter turns out to be β_1, which is associated with $\bar{n}_e$. The robust estimation technique in MATLAB also delivers good results (in fact, not much worse than GLS), as it is designed to cope with outliers. However, we will see that in the next experiment ROB does not perform well at all.

4.2. Effect of Logarithmic Transformation

We next tested the effect of a logarithmic transformation, which is often used to transform a power-law regression model into a linear form. However, the logarithm alters the data distribution, which may lead to misguided inferences from OLS [2,3]. Therefore, the flexibility offered by GLS is expected to be beneficial in this case, as it allows the observed distribution to deviate from the modeled distribution.

4.2.1. Single Predictor Variable

Again, we first performed a simple regression experiment involving a single predictor variable, with a power law deterministic model and additive Gaussian noise on all variables. In accordance with the typical situation of fitting fusion scaling laws to multi-machine data, the noise standard deviation was taken proportional to the simulated measurements, corresponding to a given set of relative error bars. As a result, in the logarithmic space the distributions were only approximately Gaussian, with the standard deviation given by the constant relative error on the original measurement (homoscedasticity). Ten points were chosen with predictor values ξ_n unevenly spread between zero and 60. A power law was proposed to relate the unobserved ξ_n and η_n:

$$\eta_n = \beta_0 \, \xi_n^{\beta_1}, \quad n = 1, \ldots, 10.$$

Then, Gaussian noise was added to the ξ_n and η_n, corresponding to a substantial relative error of 40%. We finally took the natural logarithm of all observed values x_n and y_n, enabling application of the same linear regression methods that were used in the previous experiment. In this particular experiment, we chose $\beta_0 = 0.8$ and $\beta_1 = 1.4$, but we found that other values yield similar conclusions. Again, 100 data replications were generated, allowing calculation of Monte Carlo averages.

The averages and standard deviations over all 100 runs are given in Table 2. Again, the results show that GLS is robust against the flawed model assumptions, now performing similar to TLS.

Table 2. Monte Carlo estimates of the mean and standard deviation for the parameters in a log-linear regression experiment with proportional additive noise on both variables.

Parameter	Original	GLS	OLS	MAP	TLS	ROB
β_0	0.80	**0.94 ± 0.47**	2.2 ± 2.3	3.0 ± 1.7	0.99 ± 0.70	2.72 ± 0.77
β_1	1.40	**1.39 ± 0.11**	1.19 ± 0.16	1.08 ± 0.26	1.41 ± 0.14	1.17 ± 0.11

4.2.2. Multiple Predictor Variables

In the last experiment with synthetic data, we studied the effect of a logarithmic transformation in a similar problem as the one described in Section 4.1.2, but in the case of a power law. In particular, the variable η was calculated for the same range of values of the parameters β_i as given in Equation (11), but now according to a power law:

$$\eta = \beta_0 \bar{n}_e^{\beta_1} B_t^{\beta_2} S^{\beta_3}.$$

Then, Gaussian noise was added to all variables. However, when applying the relatively low noise levels used in Section 4.1.2, no significant differences were observed in the performance of GLS and MAP. Therefore, the noise levels for the predictor variables were augmented to 20% for $\bar{n}_e$ (variable x_1), 5% for B_t (variable x_2) and 15% for S (variable x_3). The level for P_{thr} was kept at 15%, as before. This is still well within the maximum variability range that can be expected for the predictor variables in the ITPA database, as discussed in Section 3.

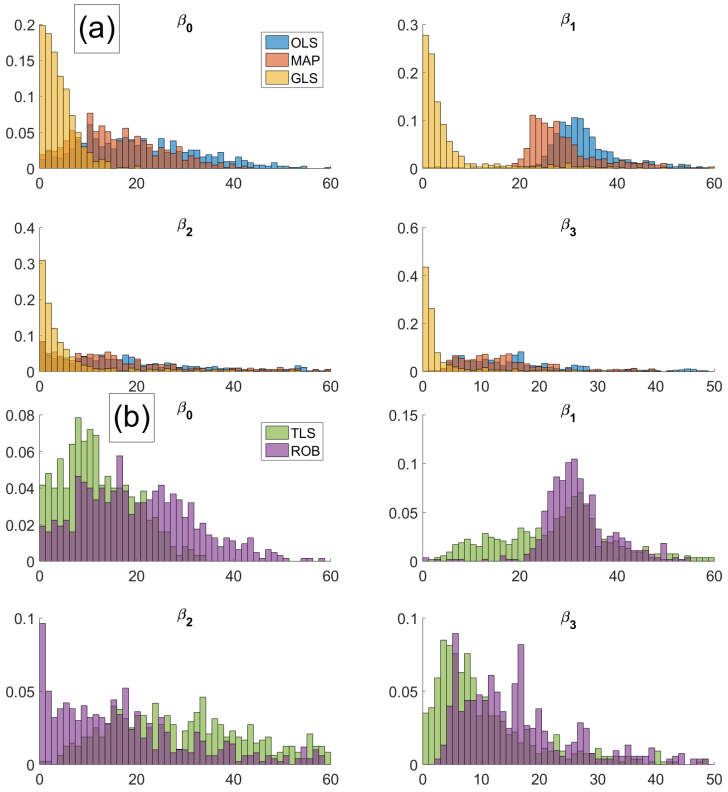

Figure 4. Histograms of the relative error in estimating the regression coefficients β_i by means of OLS, MAP and GLS for a power-law regression problem after a logarithmic transformation. Horizontal axes represent the error in percent and vertical axes probability, normalized to one. (**b**) Similar, for TLS and ROB.

After adding the noise, all data were transformed to the logarithmic domain, and 10 datasets were generated for each combination of regression coefficients. Subsequently, linear regression analysis was applied to each of the log-transformed datasets. The coefficient estimates, defined as the average over the 10 replications, were then compared among the various regression methods, as shown in Figure 4. Again, the normalized histograms of the relative error on the estimated parameters are displayed, showing

the consistently better performance of GLS over all other methods tested, including TLS and ROB. For GLS, the errors on β_0 and β_1 are the largest, compared to those on β_2 and β_3, but the majority is still below 20%. As for β_0, the slightly inferior performance of GLS relative to the results with outliers in Section 4.1.2 is simply due to the fact that $\log \beta_0$ for the lowest values of β_0 is negligibly small compared to $\log \eta - \log \beta_0$.

5. Power Threshold Scaling

We finally come to the application of power threshold scaling using real-world data from the ITPA database for all variables, including the response variable P_{thr}. We start with log-linear regression and then apply nonlinear regression analysis. Next, we perform a simple analysis of the influence of the error bars on the estimation results, and we finally provide a discussion of the results in this section.

5.1. Linear Scaling

We first followed the standard practice of transforming to the logarithmic scale to estimate the coefficients β_0, β_1, β_2 and β_3 in Equation (9) via linear regression. In the GLS method, we introduced additional parameters $\sigma_{\mathrm{obs},\alpha}$ ($\alpha = 1, \ldots, N_{\mathrm{t}}$), one for each of the $N_{\mathrm{t}} = 8$ tokamaks contributing data to the scaling. That is, if a certain data point with index n originated from tokamak α, then in term n of the objective function in Equation (8), an observed distribution was used, parameterized by means of the $\sigma_{\mathrm{obs},\alpha}$ corresponding to that machine. The $\sigma_{\mathrm{obs},\alpha}$ serve a similar purpose as the parameter σ_{obs} defined above, except that they describe the observed standard deviations of the logarithmic power threshold. This, of course, corresponds to the relative errors on the power threshold itself. To calculate σ_{mod} for each data point, we used the relative measurement error bars quoted in the database (typically 4% for $\bar{n}_{\mathrm{e}}$, 1% for B_{t}, 3% for S and 15% for P_{thr}). Considering the discussion in Section 3 regarding other sources of uncertainty, it is clear that the $\sigma_{\mathrm{obs},\alpha}$ will need to take into account other, "unexpected" uncertainty sources, hence increasing the flexibility of the method.

In this analysis, we compared GLS only with OLS and the powerful MAP method. The results on the IAEA02 data are given in Table 3. The predictions for ITER are also shown, for two typical densities (0.5 and 1.0×10^{20} m^{-3}). All estimates are accompanied by their 95% credible intervals obtained from 100 bootstrap samples (artificial datasets). We stress that this notion of a credible interval corresponds to the standard Bayesian definition of an interval wherein the true value of a stochastic variable is assumed to lie with a certain probability (e.g., 0.95).

The estimates by GLS of the parameters $\sigma_{\mathrm{obs},\alpha}$ (observed standard deviation on $\log P_{\mathrm{thr}}$), for each of the devices contributing to the IAEA02 data, were expressed as a relative error on the bootstrap-averaged P_{thr}. These relative errors and their credible intervals are given in Table 4. The relative error on the power threshold lies around 15% to 30% for the various machines, except for ASDEX, where the uncertainty reaches a higher level of about 40%. On average, this yields an estimated error of 24.2% for P_{thr}, which is quite somewhat higher than the average of 15% mentioned in the database, although still considerably lower than the upper bound of 38%, as calculated in Section 3. Again, this is an indication of additional sources of uncertainty, on top of mere measurement error, causing the data points to deviate from the

proposed regression model, as discussed already in Section 3. That extra uncertainty is captured by the GLS method.

Table 3. Estimates of regression parameters and predictions for ITER in log-transformed linear scaling of the H-mode threshold power using the IAEA02 dataset. The bootstrap averages are given, as well as the 95% credible intervals (CI).

Method		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{P}_{thr,0.5}$ (MW)	$\hat{P}_{thr,1.0}$ (MW)
OLS	Average	0.0507	0.485	0.873	0.843	38.0	53.2
	CI	±0.0060	±0.073	±0.061	±0.041	±4.4	±8.0
MAP	Average	0.0449	0.567	0.867	0.901	45.6	67.6
	CI	±0.0051	±0.078	±0.069	±0.039	±5.0	±9.6
GLS	Average	0.0426	0.660	0.795	0.946	48.3	76.4
	CI	±0.0042	±0.069	±0.059	±0.034	±4.7	±9.8

Table 4. Estimates of the observed standard deviations $\sigma_{obs,\alpha}$ of the logarithmic power threshold, expressed as percentage errors on P_{thr} itself, for the tokamaks contributing to the IAEA02 dataset, obtained using log-transformed linear scaling. The bootstrap averages are given, as well as the 95% credible intervals (CI).

	ASDEX	**AUG**	**CMOD**	**DIII-D**	**JET**	**JFT-2M**	**JT-60U**	**PBXM**
Average (%)	41.8	23.0	22.0	15.7	24.6	15.9	22.8	27.6
CI (%)	±5.3	±1.4	±1.1	±1.8	±2.0	±1.2	±2.3	±2.9

5.2. Nonlinear Scaling

Next, we show the results of nonlinear regression in the original data space, *i.e.*, without logarithmic transformation. Whereas this prevents an analytic solution using OLS, the advantage is that the distribution of the data is left undistorted [2,3], while the implementation of OLS, MAP and GLS is not significantly more complex. Indeed, the distribution of the right-hand side in Equation (9) can be approximated by a Gaussian with mean $\mu_{mod} = \beta_0 \bar{n}_e^{\beta_1} B_t^{\beta_2} S^{\beta_3}$ and standard deviation σ_{mod}, given by:

$$\sigma_{mod}^2 = \sigma_{P_{thr}}^2 + \mu_{mod}^2 \left[\beta_1^2 \left(\frac{\sigma_{\bar{n}_e}}{\bar{n}_e} \right)^2 + \beta_2^2 \left(\frac{\sigma_{B_t}}{B_t} \right)^2 + \beta_3^2 \left(\frac{\sigma_S}{S} \right)^2 \right].$$

Hence, the modeled standard deviations depend on the measurements (heteroscedasticity). Nevertheless, in defining the observed standard deviations $\sigma_{obs,\alpha}$, we introduced an approximation assuming constant error bars for all measurements from a single machine. This assumption may be relaxed in the future.

The results of the scalings and predictions are presented in Tables 5 and 6. We compared GLS with OLS and MAP using uniform priors. It may be possible to derive even less informative priors for MAP, as was done in the log-linear case in Section 5.1 (and see [30,35]), but this was not pursued here. Moreover,

even in the log-linear analysis, we observed only a marginal difference between the results under various choices of priors.

Table 5. Estimates of regression parameters and predictions for ITER in power-law scaling on the original scale of the H-mode threshold power using the IAEA02 dataset. The bootstrap averages are given, as well as the 95% credible intervals (CI).

Method		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{P}_{thr,0.5}$ **(MW)**	$\hat{P}_{thr,1.0}$ **(MW)**
OLS	Average	0.0274	0.773	0.96	1.038	69	118
	CI	±0.0083	±0.090	±0.10	±0.071	±15	±32
MAP	Average	0.0425	0.643	0.788	0.933	44.2	69.1
	CI	±0.0041	±0.074	±0.079	±0.034	±3.8	±8.2
GLS	Average	0.0397	0.715	0.751	0.984	51.6	84.7
	CI	±0.0036	±0.071	±0.081	±0.031	±4.0	±8.8

Table 6. Estimates of the observed standard deviations $\sigma_{obs,\alpha}$ of the power threshold P_{thr}, expressed as percentage errors, for the machines contributing to the IAEA02 dataset, obtained using power-law scaling. The bootstrap averages are given, as well as the 95% credible intervals (CI).

	ASDEX	**AUG**	**CMOD**	**DIII-D**	**JET**	**JFT-2M**	**JT-60U**	**PBXM**
Average (%)	35.8	21.2	20.4	15.9	22.4	15.7	22.3	27.7
CI (%)	±9.1	±4.3	±3.4	±2.4	±3.8	±2.2	±4.6	±8.1

It should also be mentioned that, in obtaining Table 6, we again calculated relative errors from the observed standard deviations estimated by GLS. However, this time, the relative errors are not the same for all measurements coming from a single machine, so we calculated an average for each machine (and similar for the credible interval). The resulting errors on P_{thr} are relatively similar to those using log-linear scaling, with an average over all devices of 22.7%, which is again higher than the 15% expected from measurement error only.

5.3. Influence of Error Bars

In the last couple of experiments, we intended to assess the sensitivity of the regression analysis on the accuracy of the error bars on the ITPA data. A systematic study of this influence is outside the scope of this paper, and as a first simple test, we doubled the error bars on all root ITPA variables (basically the electron density and the magnetic field together with various geometrical plasma parameters and sources of input power), which were used for calculation of the variables involved in the power threshold scaling law. On average, over all machines, this resulted in the following derived error bars: 9% on $\bar{n}_e$, 2% on B_t, 5% on S and 32% on P_{thr}. Again, these are all below the maxima quoted in Section 3.

We then performed power-law regression with MAP and GLS on the ITPA data using these larger error bars; the results are given in Table 7 [36]. It is observed that, based on MAP, the predictions for ITER are lowered relative to the analysis with the original error bars in Section 5.2. In contrast, the predictions by GLS remain about the same as before. On the other hand, the GLS estimates of the observed standard deviations, listed in Table 8, are increased for all devices. This is how GLS accommodates the increased error bars on the data.

Table 7. Estimates of regression parameters and predictions for ITER in power-law scaling on the original scale of the H-mode threshold power using the IAEA02 dataset with all error bars (on the root quantities) doubled.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{P}_{thr,0.5}$ **(MW)**	$\hat{P}_{thr,1.0}$ **(MW)**
MAP	0.0436	0.581	0.828	0.900	41.0	61.3
GLS	0.0393	0.725	0.742	0.990	52.1	86.2

Table 8. Estimates of the observed standard deviations $\sigma_{obs,\alpha}$ of the power threshold P_{thr}, expressed as percentage errors, for the machines contributing to the IAEA02 dataset with all error bars doubled, obtained using power-law scaling.

ASDEX	**AUG**	**CMOD**	**DIII-D**	**JET**	**JFT-2M**	**JT-60U**	**PBXM**
49.5	35.9	31.7	24.9	32.9	27.6	38.9	47.7

In another simple test, we changed the error bars on $\bar{n}_e$, B_t, S and P_{thr} to values computed from the average percentages mentioned earlier in Section 3: 4% for $\bar{n}_e$, 1% for B_t, 3% for S and 15% for P_{thr}. These are averages over all machines, rendering the final absolute error bars (standard deviations), computed from the relative errors, less precise. The estimation results using power-law regression with MAP and GLS are shown in Table 9. The results of both methods are clearly affected by the averaging step, but again, MAP is seen to be more sensitive to the change in the error bars compared to GLS, which maintains estimates in a similar range as those given in Tables 3 and 5. The estimates of the observed standard deviations, given in Table 10, are adjusted accordingly by GLS.

Table 9. Estimates of regression parameters and predictions for ITER in power-law scaling on the original scale of the H-mode threshold power using the IAEA02 dataset with averaged error bars.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{P}_{thr,0.5}$ **(MW)**	$\hat{P}_{thr,1.0}$ **(MW)**
MAP	0.0488	0.552	0.807	0.862	35.1	51.5
GLS	0.0429	0.647	0.780	0.938	45.7	71.5

Table 10. Estimates of the observed standard deviations $\sigma_{\mathrm{obs},\alpha}$ of the power threshold P_{thr}, expressed as percentage errors, for the machines contributing to the IAEA02 dataset with averaged error bars, obtained using power-law scaling.

ASDEX	AUG	CMOD	DIII-D	JET	JFT-2M	JT-60U	PBXM
29.6	19.1	20.5	19.5	22.5	18.1	18.7	20.4

5.4. Discussion

Several interesting observations can be made from the experiments regarding the power threshold scaling in this section. First, considering Tables 3 and 5, it should be noted that there are several instances where the regression parameters estimated by OLS differ significantly from those obtained by GLS. For log-linear regression, this is particularly the case for the dependence of the power threshold on density and surface area and for the predicted power thresholds for ITER, as shown by the non-overlapping credible intervals. For power-law regression, the difference is rather situated in the dependence on the magnetic field. In this case, the power thresholds predicted by OLS are also quite different from the results given by GLS, but this time, the credible intervals on the OLS estimates are so wide, that they overlap with those obtained from GLS. Apart from this discrepancy, the three methods provide comparable absolute error bars on their estimates.

Furthermore, we see that the correspondence between GLS and MAP is significantly better, although the remaining differences become particularly clear for the predicted power at higher density in ITER. The estimate by GLS is higher than that provided by MAP, especially for power-law scaling.

In addition, and quite remarkably, when comparing the coefficient estimates and predictions obtained by GLS between the linear and nonlinear case, relatively consistent results can be noted. The same goes for the MAP estimates. In contrast, OLS provides widely different results, depending on whether a linear (log-transformed) or nonlinear (power-law) model is used. The relatively good consistency of the GLS estimates across regression models is a solid argument in favor of the method.

Another noteworthy point comes from the results of the two additional tests with increased and averaged error bars. They indicate that for MAP (and maximum likelihood) regression, reliable estimates of the variability of the measurements is important. However, as discussed in Section 3, the standard error bars that were used in the analysis in Sections 5.1 and 5.2 are small compared to the actual variability of the data around the theoretical scaling law. Hence, one could speculate whether the results of the MAP analysis are in fact trustworthy, given its sensitivity to the error bars on the data. Therefore, at least for MAP, it would be better to encode the available information on the error bars in sufficiently wide prior distributions (which, incidentally, would be possible for GLS, too).

A related comment is that GLS is clearly less vulnerable to inaccurate error specification compared to MAP. The mechanism behind this behavior is similar to the one that makes GLS less sensitive to outliers, *i.e.*, the observed standard deviation is able to capture deviations from the expected data variability with respect to the model. In the simple implementation of the GLS method used in the present paper, the distinction that is made between the modeled and observed standard deviation is the main difference between GLS and MAP.

6. Conclusions

Regression and scaling laws represent crucial tools in science in general and in the analysis of complex physical systems in particular. We have presented geodesic least squares regression (GLS) as a method that is able to handle large uncertainties on the data and on the regression model, and we have demonstrated its application to power-law regression. Operating on a manifold of probability distributions, GLS has the advantage that its results can be easily visualized in the case of the univariate Gaussian distribution. However, GLS is sufficiently flexible to allow tackling much more general regression problems within the same framework.

We have shown two examples of the enhanced robustness of the method using synthetic data. GLS showed a better stability in the presence of outliers and under a logarithmic transformation of a power-law, compared to established techniques. In addition, we have addressed the scaling of the L-H power threshold in magnetically-confined fusion plasmas. On the basis of data from a multi-machine database, it was observed that geodesic least squares provides estimates of regression parameters and predictions that are consistent across different regression models, in contrast to ordinary least squares. Furthermore, because GLS allows the data uncertainty predicted by the model to be different from the empirically observed uncertainty, whereas with maximum *a posteriori* they are, by design, the same, GLS is more flexible and robust at the same time. As a consequence, the degrees of freedom provided by the parameters of the regression model better serve their actual purpose: to parameterize a model that best describes a trend in the data, with minimal distraction by the data "noise".

In future work, we intend to present a more general formulation of geodesic least squares, targeted at a wider class of regression problems. In addition, various theoretical performance issues need to be addressed, including uniqueness and convergence properties of the optimization problem, asymptotic behavior of the parameter estimates, *etc*. On the practical side, we aim at establishing a broader basis for the performance of the GLS method on simulated data. This should increase the confidence over a wider range of regression problems, as well as deviations from the regression model.

Finally, although we have noted that GLS performs regression on a probabilistic manifold, we have actually made little use of the geometrical structure of the manifold, save for calculating geodesic distances. Nowadays, there are various schemes, more sophisticated than a least-squares approach, to perform regression on manifold-valued data. From that point of view, one can expect advantages of a method performing regression between probability distributions, each of them containing more information than structureless data points in a Euclidean space. One possibility that we will explore in future work is a Bayesian regression method on a probabilistic manifold, by describing the distribution corresponding to the regression model intrinsically on the manifold [37]. At the same time, this will provide uncertainty estimates on the parameters through the posterior distribution.

Acknowledgments

The author wishes to acknowledge the ITPA Topical Groups on Transport and Confinement and on Pedestal and Edge Physics for maintaining and kindly providing the data in the H-mode threshold databases.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Doyle, E.J.; Houlberg, W.A.; Kamada, Y.; Mukhovatov, V.; Osborne, T.H.; Polevoi, A.; Bateman, G.; Connor, J.W.; Cordey, J.G.; Fujita,T.; *et al.* Chapter 2: Plasma confinement and transport. *Nucl. Fusion* **2007**, *47*, S18–S127.
2. Xiao, X.; White, E.P.; Hooten, M.B.; Durham, S.L. On the use of log-transformations *vs.* nonlinear regression for analyzing biological power laws. *Ecology* **2011**, *92*, 1887–1894.
3. McDonald, D.; Meakins, A.J.; Svensson, J.; Kirk, A.; Cordey, J.G.; ITPA H-mode Threshold Database WG. The impact of statistical models on scalings derived from multi-machine H-mode threshold experiments. *Plasma Phys. Control. Fusion* **2006**, *48*, A439–A447.
4. Verdoolaege, G. Geodesic least squares regression on information manifolds. *AIP Conf. Proc.* **2013**, *1636*, 43–48.
5. Verdoolaege, G. Geodesic least squares regression for scaling studies in magnetic confinement fusion. *AIP Conf. Proc.* **2014**, *1641*, 564–571.
6. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2011; Volume 120.
7. McCullagh, P.; Nelder, J. *Generalized Linear Models*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 1989; Volume 37.
8. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: New York, NY, USA, 2000.
9. We follow standard notational practice from differential geometry with respect to index placement in the following definitions for the metric, Christoffel symbols and geodesic distance. However, in the remainder of the paper we will revert to subscript indices only, in order to avoid other notational problems.
10. Oprea, J. *Differential Geometry and Its Applications*, 2nd ed.; The Mathematical Association of America: Washington, DC, USA, 2007.
11. Verdoolaege, G.; Scheunders, P. On the geometry of multivariate generalized Gaussian models. *J. Math. Imaging Vis.* **2011**, *43*, 180–193.
12. Kass, R.; Vos, P. *Geometrical Foundations of Asymptotic Inference*; Wiley: New York, NY, USA, 1997.
13. Verdoolaege, G.; Scheunders, P. Geodesics on the manifold of multivariate generalized Gaussian distributions with an application to multicomponent texture discrimination. *Int. J. Comput. Vis.* **2011**, *95*, 265–286.
14. Kullback, S. *Information Theory and Statistics*; Dover Publications: New York, NY, USA, 1968.
15. Atkinson, C.; Mitchell, A. Rao's distance measure. *Indian J. Stat.* **1981**, *48*, 345–365.
16. Burbea, J.; Rao, C. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multivar. Anal.* **1982**, *12*, 575–596.

17. Nielsen, F.; Nock, R. Visualizing hyperbolic Voronoi diagrams. In Proceedings of the 30th Annual Symposium on Computational Geometry (SOCG'14), Kyoto, Japan, 8–1 June 2014; p. 90.
18. Beran, R. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **1977**, *5*, 445–463.
19. Pak, R. Minimum Hellinger distance estimation in simple regression models; distribution and efficiency. *Stat. Probab. Lett.* **1996**, *26*, 263–269.
20. Rao, C. Differential metrics in probability spaces. In *Differential Geometry in Statistical Inference*; Institute of Mathematical Statistics: Hayward, CA, USA, 1987.
21. Gill, P.; Murray, W.; Wright, M. *Numerical Linear Algebra and Optimization*; Addison Wesley: Boston, MA, USA, 1991; Volume 1.
22. Casella, G.; Berger, R. *Statistical Inference*, 2nd ed.; Cengage Learning: Hampshire, UK, 2002.
23. Snipes, J.A.; Greenwald, M.; Ryter, F.; Kardaun, O.J.W.F.; Stober, J.; Valovic, M.; Valovic, S.J.; Sykes, A.; Dnestrovskij, A.; Walsh, M.; *et al.* Multi-Machine global confinement and H-mode threshold analysis. In Proceedings of the 19th IAEA Fusion Energy Conference, Lyon, France, 14–19 October 2002.
24. Martin, Y.R.; Takizuka, T.; The ITPA CDBM H-mode Threshold Database Working Group. Power requirements for accessing the H-mode in ITER. *J. Phys. Conf. Ser.* **2008**, *123*, 012033.
25. Ryter, F.; The H-Mode Database Working Group. H Mode power threshold database for ITER. *Nucl. Fusion* **1996**, *36*, 1217–1264.
26. Ryter, F.; The H-Mode Threshold Database Group. Progress of the international H-Mode power threshold database activity. *Plasma Phys. Control. Fusion* **2002**, *44*, A415–A421.
27. ITPA—Threshold database. Available online: http://efdasql.ipp.mpg.de/threshold (accessed on 30 June 2015).
28. Whereas the most recent update of the database dates from 2008 [24], we used the earlier version from 2002, because it allows a better illustration of the advantages of GLS with respect to other methods. The reason is that the data in the most recent version is significantly better conditioned, in which case even a simple regression technique such as OLS turns out to be able to provide acceptable estimates of the regression parameters. This point is not relevant for the present discussion, as here our aim is to demonstrate the advantages of GLS in cases where the data are not in the best shape.
29. Verdoolaege, G.; Karagounis, G.; Tendler, M.; van Oost, G. Pattern recognition in probability spaces for visualization and identification of plasma confinement regimes and confinement time scaling. *Plasma Phys. Control. Fusion* **2012**, *54*, 124006.
30. Preuss, R.; Dose, V. Errors in all variables. *AIP Conf. Proc.* **2005**, *803*, 448–455.
31. Markovsky, I.; van Huffel, S. Overview of total least-squares methods. *Signal Process.* **2007**, *87*, 2283–2302.
32. Maronna, R.; Martin, D.; Yohai, V. *Robust Statistics: Theory and Methods*; Wiley: New York, NY, USA, 2006.
33. *MATLAB and Statistics Toolbox Release 2015a*; The Mathworks Inc: Natick, MA, USA, 2015.

34. We use the notation η for the response variable instead of P_{thr} because in this experiment η is generated artificially and therefore it is not necessarily related to the actual power threshold in fusion devices.
35. Von Toussaint, U.; Frey, M.; Gori, S. Fitting of functions with uncertainties in dependent and independent variables. *AIP Conf. Proc.* **2009**, *1193*, 302–310.
36. OLS is not repeated here because it does not depend on the error bars.
37. Pennec, X. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vis.* **2006**, *25*, 127–154.